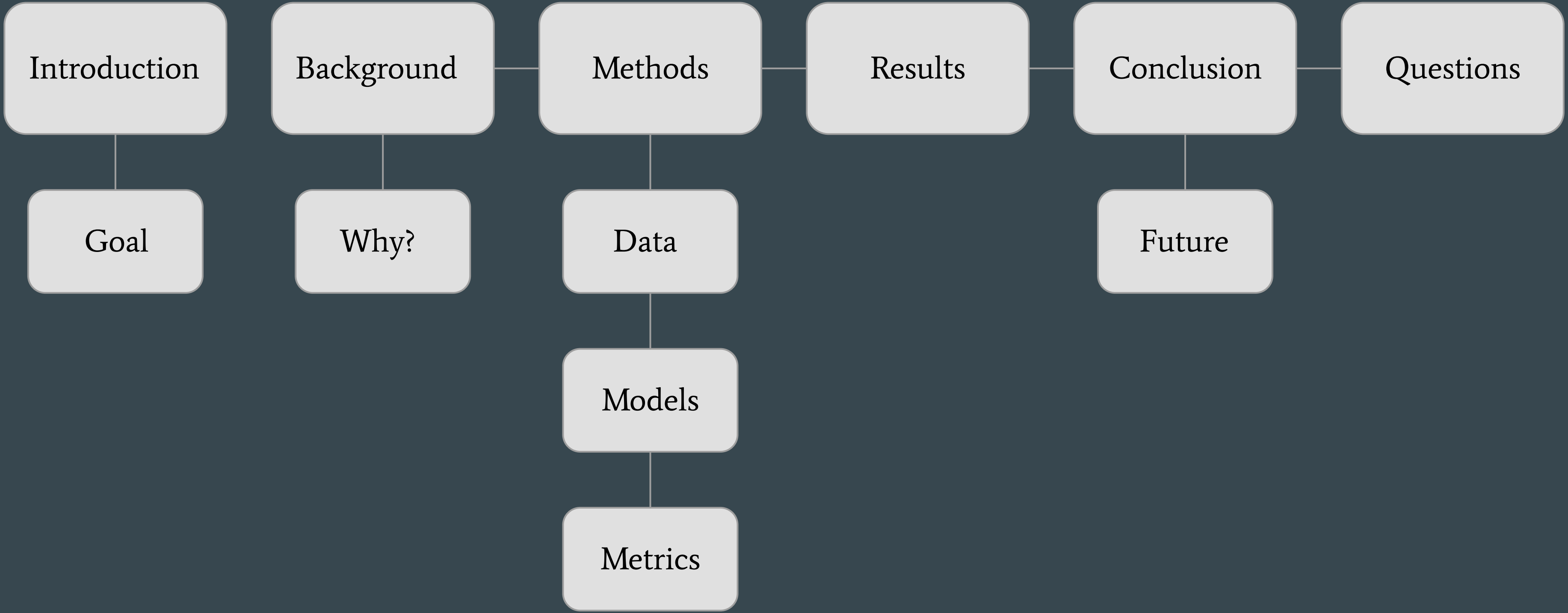
Flight Price Prediction

•••

Nicholas Gdula, Alex Mendez
CSC 642 - R

Overview



- Introduction
  - Goal
- Background
  - Why?
- Methods
  - Data
  - Models
  - Metrics
- Results
- Conclusion
  - Future
- Questions

Introduction

Goal

The purpose of this project is to predict, using multiple regression methods, the price of flights that could be invaluable to customers seeking to find the most optimal time to purchase a ticket.

Why?

Flight prices are dynamic as they are constantly changing on various factors, on one day a flight may be one hundred dollars and on the next it can double in price.

For Who?

A lot of corporations are employing data scientists to create price adjusting models that will optimize their profit margins, our tool is meant to be used by the consumer to use as a baseline for what a flight should cost under normal circumstances.

Project objective:

- For Consumers
- For Price Comparison
- For competition against industry

Understanding the price prediction climate

Current Research

Cutting edge research is diving more into Recurrent Neural Networks (RNN), as patterns in Flight data change over time.

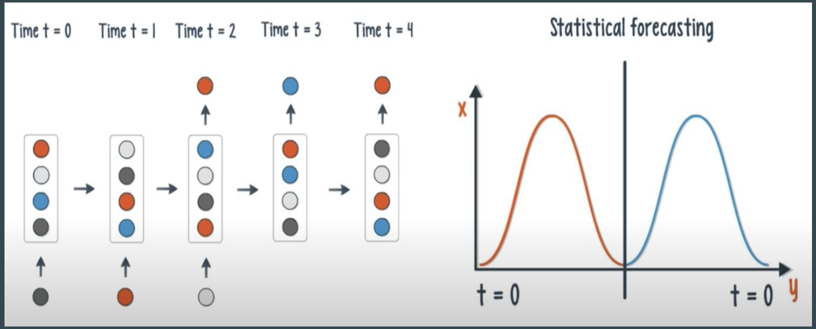
The current research:

- Common Regression Models
  - Linear Regression, RF regression, Ridge Regression, etc.
- Partial least squares approach.
- Support Vector Machines
- Linear Quantile Blended Regression

Current and Cutting Edge Research

Cutting Edge

- Biswas, Prithviraj. “Flight Price Prediction: A Case Study.” *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 6, 2022, pp. 89–94., https://doi.org/10.22214/ijraset.2022.43666.
    - Recurrent Neural Networks



Source: Deep LearningTV

Current

- William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the 2013 international conference on autonomous agents and multi-agent systems
    - Partial Least Squares
- Supriya Rajankar, Neha sakhrakar and Omprakash rajankar “Flight fare prediction using machine learning algorithms” International journal of Engineering Research and Technology (IJERT) June 2019.
    - Common ML algorithms
- T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices"
    - Linear Quantile Blended Regression

Methods (Data)

- The Dataset provided to us had three files
    - Clean dataset (https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction?select=Clean_Dataset.csv)
    - A dataset containing business class flight data
    - A dataset containing economy class flight data
- There were no missing values
- Our plans included using business and economy class flight details as a part of our model, this meant we only needed to deal with the clean dataset
    - Contained 300261 rows for all combinations of 11 predictors
    - Data collected over a fifty day period, February 11th to March 31st, 2022

Methods (Data)

- Nine variables from original dataset that were considered for use
    - Excluded Flight ID number and Flight Code
    - Airline : Airline Company (Categorical)
    - Source_city: City in which flight is departing (Categorical)
    - departure_time: Time in which flight leaves the source city
    - stops: The number of layovers for flight (Ordinal)
    - arrival_time : Time landed in destination city
    - destination_city: City in which flight is arriving (Categorical)
    - class: Ticket class (Ordinal)
    - duration: Duration of the flight from the source city to destination city
    - days_left: The amount of days remaining until flight departure
    - price: Flight Price (target value)

Methods (Data)

- Performed Exploratory Data Analysis to get a better of our variables and how we could use them
    - Reprocessing needed
- Encode ordinal variables: First encoded two ordinal variables, "stops" and "class", into numeric values by assigning them levels using the factor() function. It then subtracts 1 from each value to obtain values from 0 to n-1, where n is the number of levels.

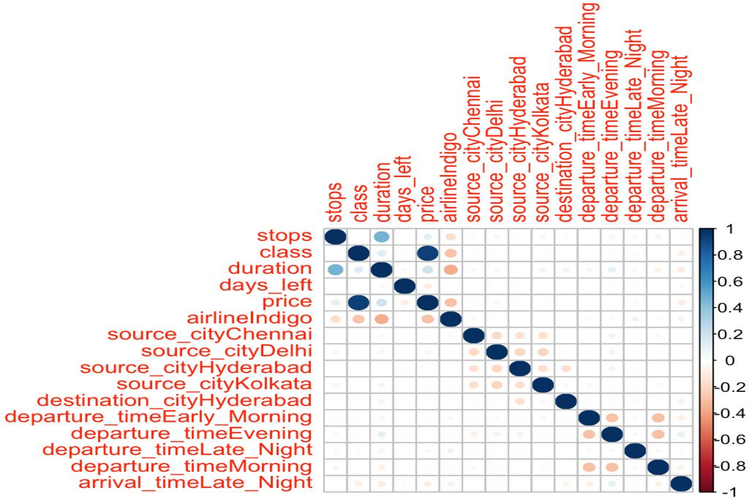
- Created dummy variables for categorical variables, "airline", "source_city", "destination_city", "departure_time", and "arrival_time" using model.matrix() function. It then replaces the original variables with the newly created dummy variables and appends them to the original data frame using cbind(). New variables were binary

- Removed Original

Methods (Data)

- This left us with 16 variables
  - stops
  - class
  - duration
  - days_left
  - airline_Indigo
  - source_cityChennai
  - source_cityDelhi
  - source_cityHyderabad
  - source_cityKolkata
  - destination_cityHyderabad
  - departure_timeEarly_Morning
  - departure_timeEvening
  - departure_timeLate_Night

- The data then was split into 70% training and 30% testing
  - 210,107 X training and 90046 for X test
  - 210,107 Y training and 90046 for Y test



Methods (Models)

The models used:

- Linear Regression
- Bagging
- Ridge Regression
- Lasso
- Random Forest Regression

Methods (Metrics)

- Data was split into training and test sets using a 70/30 split
- Create dummy variables for categorical features
- Cross validation with k=10 was used
- Adjusted R-squared and Root Mean Squared Error(RMSE) was used for evaluation of models
- Average Price of Flight was 261 US dollars

Results

	Adjusted R-Squared	RMSE (Rupees)
Linear Regression	0.904634	7005.708
Bagging	0.962823	4374.175
Ridge Regression	0.904634	7005.704
Lasso	0.904635	7005.698
Random Forest	0.964399	4280.438 ($52.14)

Conclusions

- Random forrest had the best model with highest adjusted R-squared and lowest RMSE with bagging in second
- Linear models all performed similarly with 6% lower adjust R-squared and around 2700 higher RMSE
- Most important predictors were class, airline, and stops made
  - Airline could be due to only a few airlines carrying business class in the dataset
  - Stops made could correlate to distance traveled

Conclusion (Future Work)

- Expand dataset to include other airlines and regions
- Change number of predictor variables
- Hypertune parameters for bagging and random forest
- Include other models like XGBoost, LDA, QDA
- Using separate models for business and economy flights
- With proper resources look to build RNN
    - RNN with 100 steps is equal to a 100 layer MLP

Questions?

Works Cited

- Biswas, Prithviraj. “Flight Price Prediction: A Case Study.” *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 6, 2022, pp. 89–94., https://doi.org/10.22214/ijraset.2022.43666.

- Groves, William, and Maria Gini. “On Optimizing Airline Ticket Purchase Timing.” *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 1, 2015, pp. 1–28., https://doi.org/10.1145/2733384.

- Gupta, Chitranjan Kumar, et al. “Dynamic Flight Price Prediction Using Machine Learning Algorithms.” *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2022, https://doi.org/10.1109/icac3n56670.2022.10074448.

- Jansen, Tim. “A Linear Quantile Mixed Regression Model for Prediction of Airline Ticket Prices.” *Radboud University* , 3 Apr. 2013.